

Video analytics evaluation: survey of datasets, performance metrics and approaches

Prepared by:

Dmitry O. Gorodnichy
Canada Border Services Agency
Ottawa ON Canada K1A 0L8

Robert Laganieri, Diego Macrini
University of Ottawa
School of Electrical Engineering and Computer Science
161 Louis Pasteur
Ottawa ON Canada K1N 6N5

Scientific Authority: Pierre Meunier
DRDC Centre for Security Science
613-992-0753

The scientific or technical validity of this Contract Report is entirely the responsibility of the Contractor and the contents do not necessarily have the approval or endorsement of the Department of National Defence of Canada.

Contract Report
DRDC-RDDC-2014-C248
September 2014

IMPORTANT INFORMATIVE STATEMENTS

PROVE-IT (FRiV) Pilot and Research on Operational Video-based Evaluation of Infrastructure and Technology: Face Recognition in Video project, PSTP 03-401BIOM, was supported by the Canadian Safety and Security Program (CSSP) which is led by Defence Research and Development Canada's Centre for Security Science, in partnership with Public Safety Canada. Led by Canada Border Services Agency partners included: Royal Canadian Mounted Police, Defence Research Development Canada, Canadian Air Transport Security Authority, Transport Canada, Privy Council Office; US Federal Bureau of Investigation, National Institute of Standards and Technology, UK Home Office; University of Ottawa, Université Québec (ÉTS).

The CSSP is a federally-funded program to strengthen Canada's ability to anticipate, prevent/mitigate, prepare for, respond to, and recover from natural disasters, serious accidents, crime and terrorism through the convergence of science and technology with policy, operations and intelligence.

© Her Majesty the Queen in Right of Canada, as represented by the Minister of National Defence, 2014

© Sa Majesté la Reine (en droit du Canada), telle que représentée par le ministre de la Défense nationale, 2014



Science and Engineering Directorate

Border Technology Division

Division Report: 2014-28 (TR)
July 2014

Video analytics evaluation:
survey of datasets,
performance metrics and
approaches

Diego Macrini,
Dmitry O. Gorodnichy,
Robert Laganriere



This page left intentionally blank

Abstract

This report presents a survey of evaluation datasets, metrics and practices used by the scientific community pertaining to the evaluation of video analytics in the video surveillance context. The focus of the survey is on the task of visually tracking people in video, and its application to loitering and tailgating detection. The related key results from the TRECVID video analytic evaluation conducted by the National Institute of Standards and Technology (NIST) are presented.

Keywords: video-surveillance, video analytics, person tracking, loitering detection, tailgating detection, technology readiness, performance evaluation, data-sets.

Community of Practice: Border and Transportation Security

Canada Safety and Security (CSSP) investment priorities:

1. Capability area: P1.6 – Border and critical infrastructure perimeter screening technologies/ protocols for rapidly detecting and identifying threats.
2. Specific Objectives: O1 – Enhance efficient and comprehensive screening of people and cargo (identify threats as early as possible) so as to improve the free flow of legitimate goods and travellers across borders, and to align/coordinate security systems for goods, cargo and baggage;
3. Cross-Cutting Objectives CO1 – Engage in rapid assessment, transition and deployment of innovative technologies for public safety and security practitioners to achieve specific objectives;
4. Threats/Hazards F – Major trans-border criminal activity – e.g. smuggling people/material

Acknowledgements

This work is done within the project PSTP-03-402BTS “PROVE-IT(VA)” funded by the Defence Research and Development Canada (DRDC) Centre for Security Science (CSS) Public Security Technical Program (PSTP) and in-kind contributions from the University of Ottawa by the following contributors:

1. **Dr. Dmitry Gorodnichy**, Research Scientist with the Science & Engineering Directorate, Canada Border Services Agency; Adjunct Professor at School of Electrical Engineering and Computer Science of the University of Ottawa.
2. **Dr. Robert Laganriere**, Professor of School of Electrical Engineering and Computer Science of the University of Ottawa.
3. **Dr. Diego Macrini**, Post-doctoral fellow at School of Electrical Engineering and Computer Science of the University of Ottawa.

The feedback from project partners: S. Matwin (University of Ottawa), E. Granger (École de technologie supérieure), RCMP, TC, CATSA, DRDC, UK HomeOffice, FBI is gratefully acknowledged.

Disclaimer

In no way do the results presented in this paper imply recommendation or endorsement by the Canada Border Services Agency, nor do they imply that the products and equipment identified are necessarily the best available for the purpose. The information presented in this report contains only the information available in public domain.

Release Notes

Context: This document is part of the set of reports produced for the PROVE-IT(VA) project. All PROVE-IT(VA) project reports are listed below.

1. Dmitry Gorodnichy, Jean-Philippe Bergeron, David Bissessar, Ehren Choy, Jacque Sciandra, “Video Analytics technology: the foundations, market analysis and demonstrations”, Border Technology Division, Division Report 2014-36 (TR).
2. Dmitry O. Gorodnichy, Diego Macrini, Robert Laganieri, “Video analytics evaluation: survey of datasets, performance metrics and approaches”, Border Technology Division, Division Report 2014-28 (TR).
3. D. Macrini, V. Khoshaein, G. Moradian, C. Whitten, D.O. Gorodnichy, R. Laganieri, “The Current State and TRL Assessment of People Tracking Technology for Video Surveillance applications”, Border Technology Division, Division Report 2014-14 (TR).
4. M. Lalonde, M. Derenne, L. Gagnon, D. Gorodnichy, “The Current State and TRL Assessment of Unattended and Left-Behind Object Detection Technology”, Border Technology Division, Division Report 2014-13 (TR).

Jointly with the PROVE-IT(FRiV) project (PSTP-03-401BIOM):

5. D. Bissessar, E. Choy, D. Gorodnichy, T. Mungham, “Face Recognition and Event Detection in Video: An Overview of PROVE-IT Projects (BIOM401 and BTS402)”, Border Technology Division, Division Report 2013-04 (TR).
6. D. Gorodnichy, E. Granger, J.-P. Bergeron, D. Bissessar, E. Choy, T. Mungham, R. Laganieri, S. Matwin, E. Neves C. Pagano, M. De la Torre, P. Radtke, “PROVE-IT(FRiV): framework and results”. Border Technology Division, Division Report 2013-10. Proceedings of NIST International Biometrics Performance Conference (IBPC 2014), Gaithersburg, MD, April 1-4, 2014. Online at <http://www.nist.gov/itl/iad/ig/ibpc2014.cfm>.

The PROVE-IT(VA) project took place from August 2011 till March 2013. This document was drafted and discussed with project partners in March 2013 at the Video Technology for National Security (VT4NS) forum. The final version of it was produced in July 2014.

Appendices: This report is accompanied by appendices which include the presentations related to this report at the VT4NS 2011 and VT4NS 2013 forums.

Contact: Correspondence regarding this report should be directed to DMITRY dot GORODNICHY at CBSA dot GC dot CA.

Contents

Abstract	3
Release notes	5
Table of Content	6
1 Introduction	8
2 Definitions	8
2.1 Scenes	8
2.2 Visual tracking	8
2.3 Loitering	10
2.4 Tailgating	10
2.5 Face tracking	10
3 General practices	11
3.1 Conferences, workshops and competitions	11
3.2 Video datasets and ground truth	11
3.3 Synthetic video generation (SVG)	11
3.4 ViPER Tools for video annotation and evaluation	12
4 Metrics	13
4.1 General performance ratios	13
4.2 Object count	14
4.3 Moving object tracking	15
4.4 Face tracking	16
4.5 Measures of Performance	16
5 Datasets	17
5.1 Datasets for People and Object Tracking	17
5.1.1 Pets 2007 benchmark:	17
5.1.2 The Video Surveillance Online Repository (ViSOR):	17
5.1.3 iLids	18
5.1.4 VIVA datasets:	19
5.2 Datasets for facedetection assisted tracking	20
5.2.1 PETS 2007	20
5.2.2 PETS 2009	20
5.2.3 Public iLids	20

6	Video Analytics Evaluation at TRECVID	21
6.1	Overview	21
6.1.1	TRECVID tasks	21
6.2	Interactive surveillance event detection	22
6.3	Event surveillance detection: key results	23
6.3.1	Training Techniques	23
6.3.2	Fusion Techniques	23
6.4	Carnegie Mellon University team	24
6.5	Computer Research Institute of Montreal team	25
6.6	Beijing University of Posts and Telecommunications team	25
6.6.1	Head Detection	25
6.6.2	Human Head Tracking	26
6.6.3	Trajectory Analysis and Person Runs Decision	26
6.7	Peking University team	26
6.7.1	Head and Shoulder Detection/Tracking	26
6.7.2	Tracking by Detection	27
6.7.3	Particle Filter Tracking by Detection	27
6.8	IRDS-CASIA team	28
6.9	TokyoTech and Canon team	28
6.10	Tianjin University of Technology team	29
7	References	31
7.1	Bibliography	31
7.2	Datasets	31
7.2.1	Human Activity	31
7.2.2	Face Detection and Tracking	32
7.2.3	Pedestrians Detection and Street Scenes from MIT	32
7.2.4	Person Re-identification	33
7.3	International Video Surveillance and Analytics Projects	34
8	Appendix A: University of Ottawa presentation at the VT4NS 2013 conference	36
9	Appendix B: University of Ottawa presentation at the VT4NS 2013 conference	40

1 Introduction

This survey is developed to provide the background for the video analytic evaluations conducted by CBSA and University of Ottawa within the PROVE-IT(VA) project. It introduces the terminology used within the project, overviews available datasets and defines those of them that are most suitable for the tasks analyzed of the PROVE-IT(VA) project, which are the following:

1. Unattended/left-behind baggage detection;
2. Person tracking in non-crowded and crowded environments,;
3. Person-baggage tagging (association) in crowded environments;
4. Object removal detection;
5. Loitering detection;
6. Tail-gating detection;
7. Camera tampering detection.

The survey also summarizes the key findings and results from the past TRECVID video analytic evaluation conferences organized by the National Institute of Standards and Technology (NIST). In 2013 the project team has prepared and submitted its own solution for one of the TRECVID conference challenges. The provided background therefore is important for putting in the context the findings of the current project with those obtained in the past.

2 Definitions

2.1 Scenes

A scene represents a dynamic environment which includes individuals and cameras. Each camera can observe one part of the scene. Even when there is only one camera, the scene will usually be larger than what the camera can capture. For example, in a loitering experiment an individual might leave the field of view of the camera, but will not leave the scene. Every experiment is associated with one scene.

2.2 Visual tracking

Visual tracking deals with identifying a moving target and following its movement across several video frames. In most applications, visual tracking must be done in real time (i.e. at frame rate) through a sequential analysis, within which the tracker is updated each time a new frame is received by the system.

In video surveillance, tracking is usually decomposed into the following tasks:

- **Background modeling:** building a model of the static scene. The challenge here is to build a model that is invariant to illuminations changes, which integrates background motion (such as trees moving under the wind) and which can handle slight camera motions (e.g. vibrations).
- **Foreground extraction:** identifying the moving objects in a video frame. This is usually a binary image obtained by comparing the current frame and the background model. From this binary image, connected components (or blobs), i.e. groups of connected pixels belonging to the image of a moving entity, are extracted. The difficulty is to obtain connected components that entirely encompass the silhouette of the moving object. Frequently, a moving entity will be broken up into several components or one component will include more than one moving object.
- **Descriptor computation.** In order to be tracked from one frame to the next, distinctive characteristics of the moving entities must be extracted. Object descriptors can be properties of the object shape, appearance, motion, position. These can be extracted from one frame or can be computed from the integration of several frame observations. Descriptors are used in blob association and also for object class identification (e.g. car vs. person).
- **Blob association.** To follow a moving entity from one frame to the other, its corresponding connected components observed in each frame must be linked together. Therefore, in a current frame, the detected blobs must be associated with the objects that have been tracked in the previous frames. A good tracker must take into account split events where one moving object breaks up into several blobs in the current frame and merge events in which several moving objects are combined into a single blob. Occlusion is another phenomenon that makes tracking challenging because observed moving objects can temporarily disappear in the video sequence (e.g. because it passes behind a large panel) to reappear at a later time.
- **Motion estimation.** As a moving object is observed over time, a good model of its motion (i.e. speed, direction, acceleration) can be built. The information about the objects motion can be used to identify the object class (e.g. car vs. people), to detect specific events or activities (e.g. a running person). Motion information is also very useful for blob association as it can be used to predict the next position of an observed moving entity based on previous observation.

To summarize, the following are the basic elements involved in visual tracking solutions:

1. a **frame** is one image of a video;
2. a **blob** is a contiguous set of foreground pixels;

3. a **trace** is a list of blob features across frames belonging to the same object;
4. a **target** is an object followed from image to image; it consists of a list of traces.

All these elements have an associated ID that allows to uniquely identify them.

2.3 Loitering

Loitering is one of specific events that a surveillance system might be able to detect. It is defined as a person who enters the scene and remains within the scene for more than T seconds, where T is an application-specific parameter. Loitering can also include subjects that briefly leave the field of view of the camera (for a period less than $T/2$ seconds) and then reappear again. They are in this case considered to have remained in the scene, even if they were not visible in the camera view for short period of time.

Since loitering requires measuring the time spent in the scene of every individual, it can be defined as the problem of tracking every person in a scene. The most challenging cases include large crowds, occlusions, and partial visibility due to limited field of view.

It is to be expected that some loitering cases cannot be reliably detected in complex scenes, as even human observers can have trouble tracking individuals in a crowd. Thus, part of the loitering problem is to automatically detect if a scene is too complex to track every individual. This output can be used to inform security personnel that the vision system is not able to perform all of its duties at a given time.

2.4 Tailgating

Tailgating is another event that can be detected by a human and possibly by a computer vision system. It is defined as two moving objects are being in close proximity (X pixels) within each other, over long period of time (T secs), with the approximately the same directional relationship from each other with respect to the vector of their motion (α degrees).

The key challenge in detecting this event is posed by the fact that both persons can be occluded and there also multiple persons in the view of the camera.

2.5 Face tracking

Face tracking refers to the ability of tracking humans in monitored scene by using their faces as a distinctive feature. Because of their particular structures, faces can be detected in videos. A face tracker would therefore associate a series of detected faces with the target of a tracked human. All collected faces associated with a given individual can then be sent to a face recognition system that will output the most probable identity of this person based on a database of face images captured during the training phase.

In the case of multiple person tracking, the challenge is to make sure that the right face is associated with the right individual. Any error can potentially confuse the face recognition system.

3 General practices

3.1 Conferences, workshops and competitions

The assessment of video analytics algorithms in the context of video surveillance is a difficult task. Over the past years, the vision community has organized many events where researchers can test their approaches on a common set of video data. These datasets and the results obtained by various approaches are often publicly available through the websites associated with these events.

The main events for evaluation of video analytics technologies are: PETS, AVSS, CLEAR, TRECVID, VACE, CAVIAR, iLids, FRH.

More information on these events is publicly available on Internet. The most relevant of them to the PROVE-IT(VA) project tasks are further described below.

3.2 Video datasets and ground truth

The objective unbiased testing of video analytics algorithm is a challenge. The key means of evaluating tracking algorithms is through the creation of a large, standard database of video sequences, ideally accompanied with ground truth annotations.

An ideal evaluation would be performed on a large number of real videos captured under a large variety of viewpoints, viewing conditions, illumination, scenes, activities, etc. Having ground truth for such a large dataset is ambitious, and so we complement the database with synthetic sequences that have a realistic appearance and annotations of the events of interest in them. The synthetic video generation module is described in the next section.

3.3 Synthetic video generation (SVG)

The use of realistic but synthetically created video sequences allows us to isolate variables that can have a negative impact in the tracking and recognition algorithms. The SVG module is currently setup to create indoor and outdoor scenes with varying lighting and different number of people. We have selected a train station as the indoor scene and a street intersection as the outdoor scene. The scenes and sequences are created with the ObjectVideo tool for the Halflife 2 video game engine¹.

Because the scenes are synthetic, we can make sure that all aspects but one are the same in every test scene. For example, we can change the characters that must be tracked while

¹References to be inserted here

making the new character follow the exact same trajectory as the previous character. In addition, we can ensure that the new target has the same interactions with other characters as the previous one. In other words, the repeatable scenarios can be used to isolate the nature of tracking failures.

We apply the principle of changing one scene variable at a time to: lighting, characters, and crowds. This means that all combinations of variable states are considered and captured in a corresponding video sequence.

- **Target Characters.**

The set of characters to track include people with different ethnicity and gender. Currently we have four subjects, but more can be added in the future.

- **Lighting Variations.** We consider two modes: bright and dark (or daylight and night time). These modes are used in both the indoor and the outdoor scenes.

- **Crowds.** The number of still and moving people in a scene usually has an effect on the ability to track the target subject. We created test cases with different compositions of crowds and different interactions between the target and the crowd. eg, the target is occluded by some people during his time spent in the scene. As a reference, we also produce scenes in which the target is alone. Thus, it would be possible to determine if the presence of other people has a negative effect on the tracking, or if the errors are associated to other factors.

- **Annotations.** The output of the SVG module is a set of sequences and ViPER XML annotations providing the bounding box and silhouette mask of every character in the scene at any given time.

3.4 ViPER Tools for video annotation and evaluation

The Video Performance Evaluation Resource (ViPER) is a system for evaluating the results of tracking people and detecting activities. The Video Processing Analysis Resource is a toolkit of scripts and Java programs that enables the markup of visual data ground truth, and allows the systems to evaluate how closely the result data approximate that truth.

The **ViPER Ground Truth Authoring Tool (ViPERGT)** allows frame by frame markup of video metadata stored in the Viper format. It is also useful for visualization.

The **ViPER Performance Evaluation Tool (ViPERPE)** is a command line performance evaluation tool. It offers a variety of metrics for performing comparison between video metadata files. With it, a user can select multiple metrics to compare a result data set with ground truth data. It can give precision and recall metrics, perform framebyframe and objectbased evaluations, and features a filtering mechanism to evaluate relevant subsets of the data.

4 Metrics

This section presents some of the most relevant metrics for the evaluation of video surveillance algorithms and systems. These metrics have been proposed and used in different evaluation programs and workshops.

Of particular interest is the ETISEO project which aims at evaluating the performances of video analytics algorithms. The main objective of the ETISEO study was to measure the dependencies between a video processing task and video characteristics while taking into account the scene type and the global difficulty level.

While metrics that are presented here can be used in various contexts, the focus here is on using them for the evaluation of people tracking.

4.1 General performance ratios

The performance of any detection algorithm can be evaluated from the following measures:

- TP: number of true positives i.e. the number of good detections;
- FP: number of false positives, i.e. the number of wrong detections;
- TN: number of true negatives, i.e. the number of correct nondetection (not always countable);
- FN: number of false negatives, i.e. the number of missed detections.

The total number of detection is $ND = TP + FP$ while the number of references is $NR = TP + FN$. From these measures, the following measures are used to quantify the global performance of an algorithm:

$$Precision = TP / (TP + FP); \quad (1)$$

$$Recall(orSensitivity) = TP / (TP + FN); \quad (2)$$

$$F1 (or FScore) = 2 * Precision * Recall / (Precision + Recall) \quad (3)$$

Precision and sensitivity are often two complementary metrics of the performance of an algorithm. Indeed, an algorithm often has a parameter (or a set of parameters) that controls the sensitivity of the detection. When sensitivity increases then, most of the time, the precision of the algorithm is reduced (more false negative being obtained). The F1 score is a way to combine the precision and recall results into a single metric.

Recallprecision graphs are often presented in which a curve is produced with the different parameter values for each compared methods. The closer the curve is to the upper right of the graph (where bot precision and recall are maximized), the better the algorithm is.

4.2 Object count

In a frame, an object can be located by specifying either:

- the position of its centroid;
- a bounding box;
- a bitmap image showing the object silhouette.

In order to determine if an object has been correctly detected by an algorithm, the decision can be based on the following measures:

- the distance between detected and reference centroid;
- the *Dice* coefficient measuring the similarity between sets. If Ar is the area of the reference object, Ad is the area of the detected object and Ird is the area of the intersection between the reference and detected objects, then $Dice = 2 * Ird / (Ar + Ad)$. Another variant of this metric is $Dice = Ird * Ird / (Ar * Ad)$. The *Dice* coefficient can be applied to either bounding boxes or binary map. This metric is however not a proper distance as the triangular inequality condition is not met. For this reason, the next metric should be preferred.
- the Jaccard distance uses the area of the union between the two sets Urd and is defined as $J = 1Ird/Urd$.

These measures do not take into account possible object split or merge. If a detected object is broken into small blobs, then, most probably, none of these blobs will have a sufficiently low Jaccard distance to the reference object. Another measure would then be to estimate the degree of overlap of the detected object with the reference one, that is $1Ird/Ad$. Then all detected blobs sufficient close to a reference object would be counted and this count will represent the SPLIT FACTOR. It is also to compute a SPLIT INDEX by averaging the inverse split factor, $1/SPLIT\ FACTOR$ over all reference object for which there is at least one match. The split index is then a number between 1 (no splitting) and 0 (infinity of split). All split blobs associated with a same reference object can then be grouped together and their set would be use to estimate the Jaccard distance to the reference object. Reciprocally, it will happen that a several close by detected objects merged into a single blobs. The idea would then be to measure the MERGE FACTOR by counting the number of reference objects associated to a single blob using the metric $1Ird/Ar$. If this distance is small, then the reference object can be considered to have been successfully detected. Again, the MERGE INDEX would be equal to the average of $1/MERGE\ FACTOR$.

4.3 Moving object tracking

The path that is followed by a tracked object is represented by a list of blobs grouped in traces that are in turn grouped in a target.

- A **blob** is associated with one video frame and one trace. A blob can be given, for example, as a bounding box, an ellipse, and/or a silhouette mask. The video frame ID associated with a frame can be considered as the detection time of the blob;
- A **trace** is a sequence of blobs in consecutive video frames;
- A **target** is a sequence of traces which need not be in consecutive video frames, i.e., there can be temporal gaps in which the object was not detected (for example because of occlusion).

There should not be more than one trace associated with the same target at the same time (i.e., on the same video frame).

Ideally there should be a one-to-one relationship between reference tracked objects (ground truth) and the identified targets. The global performance of the tracker can then be measured as follows:

- **TRACKING TIME**: the percentage of frames in which the reference object is being tracked by system. For each blob of the reference tracked object, it is checked if at least one target contains a detected blob matching it using one of the criteria defined above. If there is a match, then the reference object is considered to be tracked in this frame. This metric however, mainly measure the detection rate of the system rather than its ability to successfully track objects. The next two metrics try to cope with this issue.
- **PERSISTENCE INDEX**: it is measured by counting the number of reference tracked objects that are associated with each identified target. A good tracker should create targets that are each associated with a single object. This metric favors conservative trackers that creates very short targets such to make sure that none of these targets are wrongly associated with more than one object. The next metric will then counterbalance this effect.
- **CONFUSION INDEX**: it is measured by counting the number of targets that are associated with each reference object. A target is considered to be associated with a reference object if at least B_{min} blobs in it are assigned to the reference object. This metric favors very lenient trackers that create very long targets to make sure they each completely include a reference tracked object.

It is also possible to measure the frame-to-frame performance of the tracker. This is done by considering each pair of consecutive frames. For each moving object visible in these two

frames, check the matching blobs in the first frame. The counted value are:

GT: good tracking, i.e. the number of detected blobs that are linked to a blob that matches the same reference object in the second frame.

FT: false tracking, i.e. the number of detected blobs that are linked to a blob that does not match the same reference object in the second frame.

MT: miss tracking, i.e. the number of reference object pairs for which there is no matching blobs or there is blob with no link.

Finally, it can also be of interest to measure the accuracy of the tracked objects trajectories. One can measure the number of reference trajectories within a certain distance of a detected trajectories and vice versa.

4.4 Face tracking

A face tracking system combines both a face detection aspect that would be measured using criteria 5.2 and a person tracking aspect that would be measured using 5.3. Note that during the tracking, all detected faces that belong to the same target are considered to belong to the same individual. It is therefore possible to measure the PERSISTENCE INDEX and the CONFUSION index of the face identities.

4.5 Measures of Performance

Since detection system performance is a tradeoff between probability of miss vs. rate of false alarms, we use the Normalized Detection Cost Rate (NDCR) measure for evaluating system performance. NDCR is a weighted linear combination of the systems Missed Detection Probability and False Alarm Rate (measured per unit time).

$$NDCR = Pmiss + \beta * RFA \text{ where :} \quad (4)$$

$$PMiss = Nmisses / NRef, \quad (5)$$

$$RFA = NfalseAlarms / NCamHrs, \quad (6)$$

$$\beta = CostFA / (CostMissRTarget) \quad (7)$$

NDCR is normalized to have the range of [0, +8) where 0 would be for perfect performance, 1 would be the cost of a system that provides no output, and +8 is possible because false alarms are included in the measure and are unlimited.

The inclusion of decision scores in the system output permits the computation of Detection Error Tradeoff (DET) curves. DET curves plot Pmiss vs. RFA for all thresholds applied to

the systems decision scores. These plots graphically show the tradeoff between the two error types for a system.

5 Datasets

5.1 Datasets for People and Object Tracking

5.1.1 Pets 2007 benchmark:

The datasets are 8 multisensor sequences containing the following 3 scenarios (1) loitering, (2) attended luggage removal (theft), and (3) unattended luggage. This dataset is publicly available, annotated, and has been used for evaluation purposes for the last 4 years. In addition, the scenes are filmed using real cameras installed in an airport. An example image is included below.



5.1.2 The Video Surveillance Online Repository (ViSOR):

This repository contains a large number of sequences. In particular, we have selected the Outdoor Unimore set (Available at http://imagerlab.ing.unimore.it/visor/video_videosInCategory.asp?idcategory=18.) This set of sequences we obtain with outdoor cameras located in the courtyard of a college campus. Such a courtyard has a significant number of columns that briefly occlude some of people walking in the scene.

In some of the sequences, there are large crowds that would make the detection of loitering quite challenging. In particular, it can be very difficult to determine if an individual from the crowd reenters the field of view of the camera. In such cases, we expect the loitering detection algorithms to label such interval of frames as scene is too complex for full tracking. When

evaluating algorithms, we would expect a minimum number of these outputs, but we prefer them over outputs with tracking errors.



5.1.3 iLids

The iLids datasets can be ordered from <http://www.homeoffice.gov.uk/scienceresearch/hosdb/ilids/> and have the following event detection scenarios: (a) sterile zone, (b) parked vehicle, (c) abandoned baggage, and (d) doorway surveillance. In addition, there is one dataset with a multiple camera tracking scenario. All the scenarios are recorded in a real airport. The drawback of this dataset is its cost (about \$200 per dataset), the fact that training and testing data are in separate datasets (i.e., doubling the cost of buying the data), and its end user licence agreement (EULA), which must be signed by every group using it. We suggest to begging by buying four datasets: the abandoned baggage and the multiple camera tracking for both training and testing.



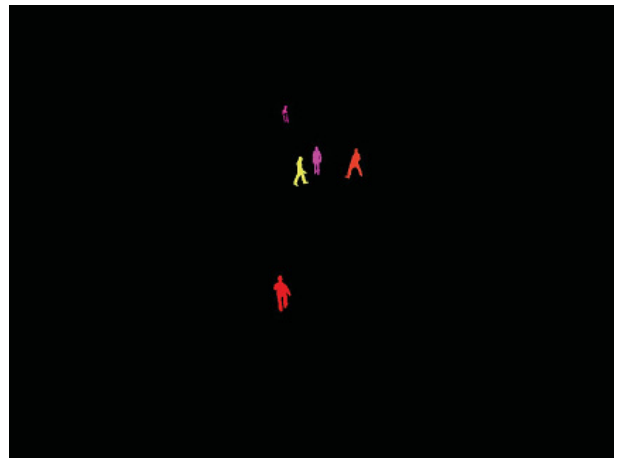


5.1.4 VIVA datasets:

We collected additional datasets to evaluate people tracking and face tracking algorithms, using one outdoor camera and one indoor camera.

For the indoor camera dataset uses 640 by 480 px video sequences that include events such as entering a door and passing by. Faces contained in a square bounding box vary in size from approximately 40 x 40 px to 125 x 125 px. In addition, we created synthetic datasets of indoor and outdoor sequences with complete and exact ground truth, including:

- 3D target centroids in the world frame;
- Per target bounding box around visible target pixels;
- Per target bounding box around all target pixels;
- Pixelwise foreground segmentation map, which assigns a label to each pixel according to the target (or background) appearing at that location.



5.2 Datasets for facedetection assisted tracking

We have evaluated several available datasets to decide which would be the best for evaluating the performance of the available tracking methods. In selecting these datasets, we kept in mind the need to test against different criteria (such as illumination changes, individuals reentering a scene, individuals grouped together) in relevant environments to this project. Most of the datasets take place in an environment similar to an airport.

5.2.1 PETS 2007

Within the PETS 2007 dataset (available at <http://pets2007.net/>), there are several views of people waiting in a crowded airport security line. We suggest using the video sequences from View 4 in this dataset. This view is approximately eyelevel, a good distance away from the line. There are several low resolution faces that would not likely be detected waiting in line, but there are several individuals and groups passing by and waiting around closer to the camera that would make for good test data. The videos in this dataset are 720 x 576 px. Natural, well focused faces in this dataset will usually have between 4 and 8 pixels between the eyes.

5.2.2 PETS 2009

Within the PETS 2009 dataset (available at <http://www.cvg.rdg.ac.uk/PETS2009/>), there are several views of sparse groups of people walking through an outdoor intersection. This dataset provides us with outdoor views in with natural illumination changes. Many of the views in this dataset are too low of resolution for face tracking, but View 8 is a good view for performing this tracking. These videos are 720 x 576 px, and well focused faces have, on average, about 4 px between the eyes.

5.2.3 Public iLids

The public iLids dataset available at http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html has two video sequences that will be good for evaluating the performance of our tracking algorithms. The Single face dataset and Multiple faces dataset are indoor videos of people moving around close to the camera to test different invariances. There are illumination shifts, people entering and leaving the scene, people occluding one another, and swapping places rapidly. These are good videos to ensure these situations are wellhandled by the algorithm being tested. These videos have a resolution of 720 x 576 px. Well focused faces in these videos have between 10 and 20 px between the eyes, the largest of these datasets (due to the artificial nature of the videos). The exception is the video titled *motinas_emilio_webcam* from the singleface dataset, which has a focused face of about 7 px between the eyes.

6 Video Analytics Evaluation at TRECVID

6.1 Overview

TRECVID is a workshop series sponsored by the US National Institute of Standards and Technology (NIST) and that has for objective to encourage the development of new technology in information retrieval in video. In order to formally evaluate the advances in the research field, TRECVID defines each year a series of video retrieval tasks and invite researchers to submit results of their approaches in solving the proposed tasks. In order to come up with a fair evaluation of the different methods, TRECVID follows a rigorous procedure that includes, for each task:

- a detailed description of the objectives of the task
- a large test collection, usually divided into a development test set and an evaluation test set.
- a complete ground truth representing the expected solution. Evaluation of the proposed solution is generally done on a set for which the ground truth is kept secret.
- a forum for discussion and presentation of the results.

A call for participation is issued each year, usually in February. The registered teams can select the tasks of their choice and then receive instructions and dataset. Results are submitted electronically and the final workshop is generally held in November of the same year. TRECVID runs each year since 2001.

6.1.1 TRECVID tasks

Each year, a variety of video retrieval tasks are proposed. Some are a repetition of the ones of the previous years while others are new. Examples of past tasks that have been abandoned since then, are:

shot boundary detection: detect in a video all the cuts and transitions that separate the different camera shots of the video.

story segmentation: given a long video, segment it into sequences of coherent stories, i.e. detects the transitions where the video switches from one topic to another one.

video copy detection: from a video query, try to find in a large video database if it exists copies of some segments of this video. The copies might have been altered by some transformations such as cropping, resizing, text overlay.

video summarization: from a long video file, create a summary containing the main elements of the original video. Typical summary duration was 30 seconds for a 30 minutes video.

In 2012, the proposed tasks are the following:

semantic indexing: automatically assign semantic tags to video segments. The task specifies a list of 500 concepts to be associated with the video segments (e.g. bicycle, telephone, flag, mountain,).

known-item search: given a textual description, the system must find the videos corresponding to this description. The description is composed of short phrases identifying what should be visible in the video.

instance search: from a query image showing an object of interest search in a video collection every instances of this object.

multimedia event detection: from a set of 20 predefined events, find a video collection all the occurrences of each of the selected events. Examples of events are birthday party, parade, dog show, etc.

multimedia event recounting: given an event description and some examples, review a video clip containing this event and produce a textual recounting that summarizes the key evidences for this event.

interactive surveillance event detection: given a search request for a particular event, produce a system that will assist a human operator to identify occurrences of this event in a set of surveillance videos. As this particular task is of interest to the context of video surveillance, more details will be given in the next section.

6.2 Interactive surveillance event detection

This task is described as follows on the NIST web site:

Given a collection of surveillance data files (e.g. that from an airport, or commercial establishment) for preprocessing, at test time take a small set of topics (search requests for known events) and for each return the elapsed search time and a list of video segments within the surveillance data files, ranked by likelihood of meeting the need described in the topic. Each search for an event by a searcher can take no more than 25 elapsed minutes, measured from the time the searcher is given the event to look for until the time the result set is considered final.

The interactive aspect of this task is new for this year. It is therefore a pilot task that for the first time will involve a human in the detection process. In previous years, a similar task was proposed but in which all event detection were performed by an autonomous computer system. Approaches used by the participants of previous years will have to be explored in order to produce results that will be exploited by the user during the interactive search. The list of proposed events to be detected in this task is:

- a person embracing another one
- a person bringing his/her cell to ear
- group of people meeting
- a person running
- an object put

- people splitting up
- a person pointing
- a person taking a picture

These events must be detected in a large collection of surveillance videos captured at the Gatwick airport in London UK. A development set is given in which ground truth annotation is available specifying where in the videos these events are occurring.

Since the approaches developed in previous TRECVID will play an important for this task, the rest of this document will summarize the most successful solutions that have been proposed.

6.3 Event surveillance detection: key results

In general, approaches to event detection use machine learning techniques involving a training phase and a fusion of the detector outputs.

6.3.1 Training Techniques

- 1. Bagging:** The basic idea of Bagging is to train multiple classifiers. The training samples for each classifier are generated by uniformly sampling with replacement. The final prediction is the combination by average the multiple classifiers.
- 2. AdaBoost:** The basic idea of AdaBoost is to train a sequence of weak classifiers by maintaining set of weights over training samples and adaptively updating these weights
- 3. Sequential Boosting:** after each Boosting iteration: the samples that are misclassified gain weight while the samples that are classified correctly lose weight. Therefore, the future weak classifier will be forced to focus on the hard samples. Finally, the combination of these weak classifiers will be a strong classifier.

6.3.2 Fusion Techniques

Early Fusion is a combination scheme that runs before classification. Both feature fusion and kernel space fusion are example of early fusion. The main advantage of early fusion is that only one learning phase is required. Two early fusion strategies, i.e., rule-based combination and multiple kernel learning, have been tried to combine kernels from different features. For rule-based combination, we use the average of the kernel matrix. Multiple kernel learning is a natural extension of average combination. It aims to automatically learn the weights for different kernel matrix. However, our experimental results show that the performance of multiple kernel learning is only slightly better than average combination. Considering that average combination is much less time consuming than multiple kernel learning, average combination is used as our early fusion method for final submission.

Late fusion happens after classification. While late fusion is easier to perform, in general, it needs more computational effort and has potential to lose the correlation in mixed feature space. Normally, another learning procedure is needed to combine these outputs, but in general, because of the overfitting problem, simply averaging the output scores together yields better or at least comparable results than training another classifier for fusion. Compared to early fusion, late fusion is more robust to features that have negative influence. In our final submission, we use both average combination and logistic regression to combine the outputs of different classifiers.

Double fusion combines early fusion and late fusion together. Specifically, for early fusion, multiple subsets of single features are fused by using standard early fusion technologies; for late fusion, we combine output of classifiers trained from single and combined features. By using this scheme, we can freely combine different early fusion and late fusion techniques, and get benefits of both. Our results show that double fusion is consistently better, or at least comparable than early fusion and late fusion.

The presentation of the most prominent past TRECVID submissions for event detection follows next.

6.4 Carnegie Mellon University team

Their approach is based on the so-called MoSIFT features. These are a variant of the well-known SIFT features extended to temporal sequences. MoSIFT extracts feature points on the moving objects and characterize them using appearance (histogram of oriented gradient) and motion information (optical flow vectors). Visual vocabularies are generated from cluster centers of MoSIFT features, which were sampled from the event video clips. The spatial distribution of actions was estimated by over-generated person detection and background subtraction. Each frame is divided into a set of non-overlapping rectangular tiles. The resulting BoW features are derived by concatenating the BoW features captured in each tile.

For training, in addition to the traditional SVM, they propose a Sequential Boosting SVM classifier to deal with the large-scale unbalanced classification problem. Their experimental results demonstrated that double fusion is consistently better than or at least comparable to early fusion and late fusion.

MoSIFT feature do a great job in human behavior representation for human action recognition, but the MoSIFT interesting points can be caused by humans, motion, light shaking, and shadows. Sampling the MoSIFT points from human body or action region, can reduce more noise interesting points.

Person detection and background subtraction methods were used to create hot regions. The hot regions were used in feature selection and building spatial Bag of Words (BoW).

6.5 Computer Research Institute of Montreal team

For the Surveillance Event Detection (SED) task, the system is based on an action recognition approach which mines spatio-temporal corners in order to detect configurations, called Compound Features, typical of an action of interest. The final detection is based on blobs around local frame-to-frame changes that are containing enough relevant compound features. They build an overcomplete set of Harris corners at various spatial scale and in the temporal domain and group corners within a 3x3x3 neighbourhood to form CF. CF are then encoded using information about cell position, scale and corner type to form transactions (or itemsets). A data mining algorithm is applied in order to extract frequent itemsets. An important issue is how to take a reliable decision in the presence of an event in a scene where many other actions are taking place (e.g. people walking). The original method derives a probability map from the firing rules as shown below, in order to improve the map, a Gaussian and temporal filtering was applied ; however this map is still difficult to threshold. Instead, a simpler approach was adopted where they segment the frame-to-frame difference image leading to areas of motions. They then compute the density of rules firing within each region of changes. The region with the higher rule density value is chosen and triggers an event if the average confidence value within this region is above a given threshold.

6.6 Beijing University of Posts and Telecommunications team

For the Event Detection tasks, Beijing University of Posts and Telecommunications made attempts at solving the sequential events of a person running, people meeting, people separating, a person placing an object, people embracing, and a person pointing. We outline the three techniques they used to solve these problems.

6.6.1 Head Detection

A common subroutine throughout all attempted approaches by Beijing University of Posts and Telecommunications is that of detecting human heads in a video sequence. The goal is to be more robust than simple face detection, so as to be able to detect heads at different rotations and orientations, rather than a head with an unoccluded forward-facing face. The process of detecting heads is done as follows:

- Perform background subtraction on the video sequence, consider just the foreground
- Compute the gradient of each foreground pixel between frames
- If a pixel has a gradient that is almost vertical, mark that pixel as a possible head-top point
- From each head-top, point, extract a ROI (region of interest)
- From this ROI, extract the HSV histogram and a histogram of oriented gradients (HOG).
- Finally, a series of SVM classifiers are created with these HSV and HOG features as the determining features for whether the query ROI is a human head.

6.6.2 Human Head Tracking

To track human heads, a tracking-by-detection approach is used.

- In each subsequent frame, detect new human head regions
- Compare these regions with the heads from the previous frame
- Compare the features of each new head with the features of the previous frames heads to get the correct track.
- Search within some allocated search radius
- Store the movements to build a trajectory for each head

6.6.3 Trajectory Analysis and Person Runs Decision

The trajectory information is extracted directly from the human-head tracking. We can trivially compute the relative speed, distance, and acceleration from the tracked trajectory. A linear combination of these gives us a score, that we will use to decide whether a person is running. Beijing University of Posts and Telecommunications originally attempted to tackle this problem via tracking-based methods, but they found that, in such cluster scenes, using tracking-based methods to detect multiple person events is hopeless.

As an alternative, a bag-of-words model on spatiotemporal features was used. MoSIFT features represent a large video sequence in an elegant way, and were extracted as the low-level features for the classification task. With these features, a bag of words model is adopted to locate the meaningful feature centres, which avoids the problem of feature divergence.

An SVM-HMM classifier is used to classify these feature points over a (spatiotemporal) sliding window. The SVM-HMM classifier will decide whether a video subsequence contains the People Meet or People Split Up events.

6.7 Peking University team

For the Event Detection tasks, Peking University has done work on pairwise events that explore the relationship between two people (People Meet, People Split Up, Embrace) and singular action events (Object Put, Pointing). Their submission is an improvement on work done in previous years for Trecvid. Their submission includes three separate versions of results, which differentiate amongst themselves by different algorithms for human detection, different algorithms for tracking, and different event detection modules.

6.7.1 Head and Shoulder Detection/Tracking

Since the video data in the Trecvid corpus consists of a many occlusions, head and shoulder detections are preferred over full human body detection.

Head and shoulder detection is handled by a linear SVM classifier. With several head and shoulder training examples, HOG (histogram of oriented gradients) features are extracted and

fed as input to the SVM. From this trained classifier, any given region can be classified in a binary fashion, is head and shoulders or is not head and shoulders. 5000 labeled images of head and shoulders were used as positive training examples, while hundreds of negative instances were also fed into the SVM to improve its discriminativity.

Testing every possible region in an image against this SVM would be costly. Instead, for the initial detection, the foreground is extracted from the image, leaving a smaller space for potential positive classifications. Furthermore, the camera's intrinsic parameters are used to approximate the possible size of a person, allowing for even more possible regions to be discarded. This leaves us with a much smaller space of potential head and shoulders.

6.7.2 Tracking by Detection

Tracking is done in a tracking-by-detection fashion. In subsequent frames, head and shoulder regions are detected and compared with the previous frame. To link two detections together to form a track, their HOG features are compared, which compares the visual appearance similarity between the two detections. On top of this, the distance from the previous frame and the scale from the previous frame are also taken into consideration. A linear combination of these three statistics makes up the probability that a new detection is a continuation of a given detection from a previous frame.

6.7.3 Particle Filter Tracking by Detection

An alternative tracking approach which uses a particle filter is explored. In this approach, the distribution of each target state is estimated by a particle filter. A constant velocity motion model is allocated to each particle. Then, that particle's weight (which is used for resampling) is computed by estimating the likelihood for each particle. The likelihood is computed as a combination from different sources: the associated detection score, the preliminary results of the detection-by-tracking algorithm mentioned above, and the classifier outputs.

For the classifiers, online Multiple Instance Learning is used. For each classifier, weak learners are selected by MIL Boost. Alternative Head-Shoulder Detection with Gradient Tree Boosting and Tracking by Multiple Hypothesis Tracking. A second approach is described for head and shoulder detection/tracking. In this approach, Gradient Tree Boosting is used to detect objects with both high accuracy and speed.

The detection algorithm is the following.

- Several positive and negative examples are taken to train the classifier
- HOG features are extracted from each of these images, to be used as the image representation
- A Cascade Gradient Boosting Tree is trained with these HOG features.
- Finally, for any new query instance, extract the HOG features and ask the Cascade

- Gradient Boosting Tree whether it is a head-shoulder pair or not.

Multiple Hypothesis Tracking is used to track multiple objects in a video. MHT uses statistical data association to deal with many common problems in tracking, such as track initiation, track termination, and track continuation. The head-shoulder detection algorithms are incorporated with MHT to construct one integrated detector/tracker.

6.8 IRDS-CASIA team

IRDS-CASIA proposed a method to solve a class of action detection problems. The actions recognized but the proposed system include CellToEar, Embrace, ObjectPut, PeopleMeet, PeopleSplitUp, PersonRuns, Pointing. The solution consists of two parts:

- 1) Feature extraction
- 2) Action classification

SFA features. - While training the system, the start frame and the end frame along with region in the image where the event had taken place are required. Matimatically SFA is defined as following: Given an I-dimensional input signal $x(t)$ where t is a time unit. SFA finds out an input-output function $g(x)$ so that the J-dimensional output signal $y(t)$ varies as slow as possible.

ASD feature - In the SFA learning, cuboids are derived from d successive frames. Thus we compute a statistical feature from d frames to represent an action sequence. SFA minimizes the average squared derivative, so the fitting degree of a cuboid to a certain slow feature function can be measured by the squared derivative of the transformed cuboid. If the value is small, the cuboid fits the slow feature function. Otherwise, the cuboid does not fit the function. Then by accumulating all the squared derivatives over all cuboids we can form the ASD feature.

6.9 TokyoTech and Canon team

TokyoTech describe a solution for surveillance event detection. The target events are PersonRuns, PeopleMeet, PeopleSplitUp. A particle filter is used to detect persons in a given scene. Once is person is detected, the system tracks the person in consecutive frames and obtains his/her trajectory. An event-trajectory model is used to detect each of the events. In order to determine the existence of a person, HOG features are extracted from a region. LBP histograms are used to associate the same person in consecutive frames. NHK Science and Technology Research Laboratories team

Members of the NHK Science and Technology Research Laboratories participated in the tasks at TRECVID 2011: a surveillance event detection task and a semantic indexing task. For surveillance event detection they proposed system consists of four steps.

In step 1, the system detects key points and tracks them until they disappear. It does this

by first extracting the foreground from an entire image by statistic background subtraction, which uses the average and variance of each pixel value. Next, it detects key points in the foreground with a Harris operator and tracks them by calculating the optical flow using the Lucas-Kanade method. The optical flows are then connected and stacked into key-point trajectories.

In step 2, extracted trajectories are segmented with reference to each distance. a dendrogram algorithm is used to cluster the trajectories. After this step, individual recognition processes are performed for each trajectory cluster.

In step 3, the system extracts features from each trajectory and describes them as a fixed dimensional histogram. It creates two types of histogram, one of which is related to the key points motion and the other to its appearance. The number of bins in both histograms is fixed, so it can describe any trajectory as a fixed dimension regardless of its length.

Finally, the action is classified in step 4. They used the bag-of-features approach and a support vector machine (SVM) to classify human actions. The two histograms created in step 3 are used as a feature for the bag-of-features approach. The SVM classifier was learned in the training phase using the trajectories around a person who had performed a target action. They originally annotated correct data using the development dataset cross-referenced with NIST annotated correct data and manually annotated bounding boxes that each contained one target person. The trajectories in the boxes were then used as correct trajectories.

6.10 Tianjin University of Technology team

Tianjin University of Technology has modeled human behaviour with the philosophy of bag of spatiotemporal feature (BoSTF) for individual event. The spatio-temporal interest points are detected for each temporal sliding window in the video sequence, and formulated. Then the extracted spatio-temporal interest points are clustered into visual keywords and hierarchical SVM classifier is used for semantic event modeling. Two major computations are applied to detect interest points:

1. SIFT point detection
2. Optical flow computation matching the scale of the SIFT points

The algorithm process a pair of video frames to find spatio-temporal interest points at multiple scales. Since SIFT was designed to detect distinctive interest points in still images so the detected points are distinctive in appearance, but they are independent of the motions in the video. For example, a cluttered background produces interest points unrelated to human actions. Clearly, only interest points with sufficient motion provide the necessary information for action recognition. Multiple-scale optical flows are calculated according to the SIFT scales. Then, as long as the amount of movement is suitable, the candidate interest point contains are retained as a motion interest point. After getting the spatio-temporal interest points, we need

describe these points. Since an action is only represented by a set of spatio-temporal point descriptors, the descriptor features critically determine the information available for recognition. Optical flow has the same properties as appearance gradients, so the same aggregation can be applied to both of them.

The main difference between optical flow and appearance description is in the dominant orientation. For human activity recognition, rotation invariance of appearance remains important due to varying view angles and deformations. Since videos are captured by stationary cameras, the direction of movement is an important (non-invariant) vector to help recognize an action. Therefore, the method omits adjusting for orientation invariance in the motion descriptors. Fig.2 shows the results of the spatio-temporal interest point detector in a Gatwick video key frame. It shows that the spatio-temporal features are able to clearly focus on areas with human activity.

7 References

7.1 Bibliography

1. Kevin Cannons. A Review of Visual Tracking. Technical Report CSE200807. York University.
2. Silogic Inria. Internal Technical note Metrics Definition. 2006.
3. Nghiem A.T., F. Bremond, M. Thonnat and V. Valentin. ETISEO. Performance evaluation for video surveillance systems. Proceedings of AVSS 2007, September, 2007, London, UK.
4. AnhTuan NGHIEM, Francois BREMOND, Monique THONNAT, Ruihua MA. A New Evaluation Approach for Video Processing Algorithms. IEEE Workshop on Motion and Video Computing, 2007.
5. Paul Over and George Awad and Martial Michel and Jonathan Fiscus and Wessel Kraaij and Alan F. Smeaton and Georges Quenot, An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics, Proceedings of TRECVID 2011, NIST, USA, <http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/tv11overview.pdf>

7.2 Datasets

7.2.1 Human Activity

1. i-LIDS datasets: UK Government benchmark datasets for automated surveillance. Surveillance Performance Evaluation Initiative (SPEVI)
2. PETS 2010 Benchmark Data: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance.
3. TREC Video Retrieval Evaluation
4. CANTATA project datasets: Datasets from past PETS workshops and many other sources. WallFlower dataset: For evaluating background modelling algorithms. Ground-truth foreground provided.
5. VISOR: Video Surveillance Online Repository: Lots of videos and ground truth.
6. Advanced Video and Signal based Surveillance: a variety datasets for tracking and detection. CAVIAR surveillance Dataset
7. MuHAVi: Multicamera Human Action Video DataA large body of human action video data using 8 cameras. Includes manually annotated silhouette data.

8. Anton Andriyenko crowd videos: videos of multiple people walking indoors.
9. Colour video and Thermal infrared datasets: Dataset of videos in colour and thermal infrared. Videos are aligned temporally and spatially. Ground-truth for object tracking is provided. INRIA Datasets: Cars, people, horses, human actions, etc.
10. BEHAVE Interactions Test Case Scenarios various scenario's of people acting out 10 types of group interactions: InGroup, Approach, WalkTogether, Split, Ignore, Following, Chase, Fight, RunTogether and Meet.
11. BEHAVE Crowd Sequence Data

7.2.2 Face Detection and Tracking

1. MIT Face Dataset
2. Videos for Head Tracking
3. Experiments on skin region detection and tracking: it includes a ground-truthed dataset CMU pose, illumination and expression (PIE) database A database of 41,368 images of 68 people. Each person imaged under 13 different poses, 43 different illumination conditions and with 4 different expressions.
4. Video database of moving faces and people. Univ. of Texas (Dallas) database for testing algorithms for face and person recognition, head/eye tracking, and computer graphics modeling of natural human motions. For each person there are nine static facial mug shots and a series of video streams. Complete data sets are available for 284 subjects and duplicate data sets, taken subsequent to the original set, are available for 229 subjects.

7.2.3 Pedestrians Detection and Street Scenes from MIT

1. MIT Street Scenes: CBCL StreetScenes Challenge Framework is a collection of images, annotations, software and performance measures for object detection [cars, pedestrians, bicycles, buildings, trees, skies, roads, sidewalks, and stores]
2. Daimler pedestrian benchmarks Large datasets (many thousands of pedestrian labels) for benchmarking pedestrian detection, classification and tracking algorithms.
3. Caltech pedestrian dataset Large dataset (many thousands of pedestrian labels) for benchmarking pedestrian detection, classification and tracking algorithms).

7.2.4 Person Re-identification

1. ViPER dataset: images of people from two different camera views, only one image of each person per camera. 45 degree angles 8 same viewpoint angle pairs or 28 different viewpoint angle pairs. 632 pedestrians image pairs taken by two different cameras. most challenging datasets for automated person Re-ID.
2. ETHZ dataset :
images of people taken by a moving camera, range of variations in person appearances. Sequence 1 with 83 pedestrians. Sequence 2 with 35 pedestrians. Sequence 3 with 28 pedestrians.
3. i-LIDS multi- camera tracking scenario (i-LIDS MCTS): i-LIDS MCTS dataset used for tracking crowded public spaces. 476 images of 119 pedestrians taken from two non-overlapping cameras, average of 4 images of each pedestrian and a minimum of 2 images, considerable illumination variations and occlusions across the two cameras.
4. CAVIAR4REID : Extract from multi-camera tracking dataset , indoor shopping mall with two cameras with overlapping views, multiple images of 72 pedestrians, 50 appear in both cameras, 22 come from the same camera,
5. i-LIDS MA and AA from the i-LIDS MCTS dataset –
iLIDS-MA: images of 40 pedestrians taken from two different cameras. 46 manually annotated images of each pedestrian are extracted from each camera. 3680 images of slightly different sizes.
iLIDS-AA: images of 100 pedestrians taken from two different cameras. automatically detected and tracked images of each pedestrian from each camera. 10,754 images of slightly different sizes. more challenges due to the possibility of errors coming from automated detection and tracking
6. V47 : videos of 47 pedestrians captured using two cameras in an indoor setting, two different views of each person (person walking in two different directions), foreground masks for every few frames of each video
7. QMUL underGround Re-IDentification (GRID) Dataset: 8 cameras with non-overlapping FOVs in a underground train station, low resolution ,significant illumination variations. 250 pedestrian images that appear in two different camera views 775 images of people in a single view.
8. SARC 3D short video clips of 50 persons, 4 predefined viewpoints captured with calibrated cameras. Useful to generate 3D body model, manually selected four frames

for each clip corresponding to predefined positions and postures of the people. 200 snapshots in total <http://imagelab.ing.unimore.it/visor/sarc3d.asp>

9. 3DPES (3D People Dataset for Surveillance and Forensics: <http://imagelab.ing.unimore.it/visor/3dpes.asp>)
8 camera non-overlapped field of views. Data during several days. People with strong variation of light condition, uncompressed images with a resolution of 704x576 pixels. Different cameras position, orientation, zoom levels. +100 persons detected more than one time in different cameras. people re-identification, people detection, tracking, action analysis and trajectory analysis.
10. RGB-D person Re-ID dataset - evaluation of depth-based features for Re-ID : depth information for each pedestrian using the Kinect, indoor scenario, Collaborative” , 79 people with a frontal view, walking slowly, no occlusions and stretched arms , ”Walking1” - ”Walking2” — same 79 people walking normally while entering the lab, ”Backwards” — back view recording of the people walking away from the lab. no guarantee that visual aspects like clothing or accessories will be kept constant. for each person: 1) a set of 5 RGB images, 2) the foreground masks, 3) the skeletons, 4) the 3d mesh (ply), 5) the estimated floor.

7.3 International Video Surveillance and Analytics Projects

1. CASSANDRA: Context Aware SenSing for Agression Detection and Risk Assessment, 2006-2009
2. CAVIAR: Context Aware Vision using Image-based Active Recognition, 2002-2005.
3. ADVISOR: Annotated Digital Video for Intelligent Surveillance and Optimized Retrieval, 2000-2003
4. AVSS: http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html
5. CANTATA: http://www.hitech-projects.com/euprojects/cantata/datasets_cantata/dataset.html
6. ADVISE: Advanced Video Surveillance archives search Engine for security applications (<http://www.advise-project.eu>)
7. VideoSense: European Centre of Excellence in Surveillance Video Analytics, Architectural and Operational Privacy Protection within an Inter-disciplinary Research Framework, 2011-now, (<http://www.videosense.eu>)
8. VANAHEIM: - Video/Audio Networked surveillance system enhancement through Human-centered adaptive Monitoring (EC-funded project which aims at studying innovative audio/video surveillance components for underground stations, www.vanaheim-project.eu)

9. BEAT: Biometrics Evaluation and Testing (BEAT) is a project funded by the European Commission, under the Seventh Framework Programme. Evaluation of identification technologies, including Biometrics (<http://www.beat-eu.org>)
10. SUBITO: <http://www.subito-project.eu/Home.htm>
11. SAMURAI: <http://www.samurai-eu.org/Home.htm>
12. ISCAPS: <http://www.iscaps.reading.ac.uk/home.htm>

8 Appendix A: University of Ottawa presentation at the VT4NS 2011 conference

PROVE-IT

Datasets and Challenges

VIVA Lab

Participation in TRECVID 2012

Interactive surveillance event detection pilot:

The use case addressed by this task is the retrospective exploration of surveillance video archives using a system designed to support the **optimal division of labor between a human user and the software** - an interactive system.

System task:

Given a collection of surveillance data files (e.g. that from an airport, or commercial establishment) for **preprocessing**, at test time take a small set of topics (search requests for known events) and for each return the elapsed search time and a list of video segments within the surveillance data files, ranked by likelihood of meeting the need described in the topic. Whether there will be a limit on each search's elapsed time is yet to be determined.

Data:

The test data will be the same as was used in the SED task in 2011.

Submissions:

Participants must submit at least one **interactive run**. An automatic version of each interactive run for comparison may also be submitted.

Evaluation:

In this pilot, it is assumed the user(s) are system experts and no attempt will be made to separate the contribution of the user and the system. **The results for each system+user will be evaluated** by NIST as to effectiveness - using standard search measures (e.g., precision, recall, average precision)- self-reported speed, and user satisfaction (for interactive runs).

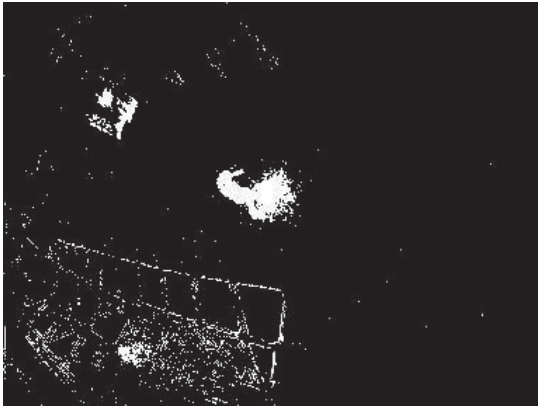
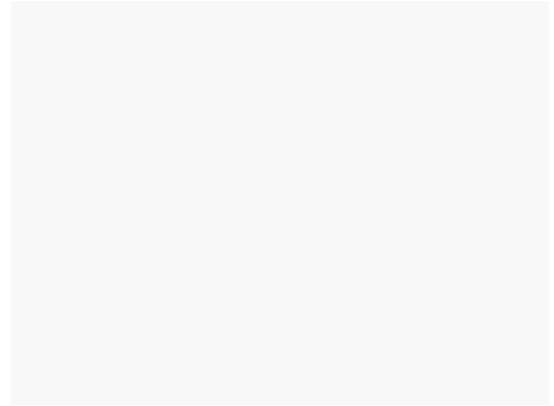
Loitering

- Challenge: Persistence of still objects.
- A large number of approaches attempt to adapt to the background, but...
- Most adaptive background subtraction algorithms will make still object “disappear” after some time.
- We should consider evaluating the state of the art in background subtraction techniques.



Novel background Generation Frameworks

- Many background subtraction techniques take any of the static objects as background. That's why they fail in the luggage-left-behind task, and other types of critical motion detection problems.
- **True Background**
Different definition from the “background” in traditional background modeling techniques such as Gaussian Mixture Modeling (GMM) and others.
- **Nonparametric Online Background Generation for Surveillance Video** (Queen's U)



Conclusions

- Consider the evaluation of background generation and subtraction algorithms.
- The identification of background is a key component in almost all tracking approaches.
- The correct detection of loitering demands the the continuous detection of still people.
- There are conference venues and workshops dedicated to this topic that we can follow.

Change Detection

IEEE Workshop on Change Detection
in conjunction with CVPR 2012



<http://www.changedetection.net/>

The detection of change, and in particular motion, in the field of view of a camera is a fundamental pre-processing step in computer vision and video processing. To date, many change detection algorithms have been developed that perform well in some types of videos but not in others. No single algorithm today seems to be able to simultaneously address all the key challenges that accompany real-world (non-synthetic) videos. This is due, in part, to the absence of a single realistic large-scale dataset with accurate ground truth providing a balanced coverage of the range of challenges present in the real world.

	Description	Scene Type	Annotated	Individuals / Crowds	Multi-camera for same scene	Lighting variations	Indoor / Outdoor	Loitering scenes	Attended luggage removal (theft) scenes	Unattended luggage scenes	Resolution and FPS	Camera types	Link
PETS 2004: CAVIAR	People walking alone, meeting with others, window shopping, fighting, passing out, and leaving a package behind.	Airport-like	yes	both	no	no	indoor	no	no	yes	384 x 288, 25 fps	wide angle	http://www.cvg.rdg.ac.uk/PETS2006/data.html
PETS 2006	Surveillance of public spaces, detection of left luggage events.	Airport-like	yes	both		no	indoor					regular	http://pets2007.net/
PETS 2007	Multisensor sequences containing scenes of loitering, attended luggage removal (theft) and unattended luggage.	Airport-like	yes	both	yes	yes	indoor	yes	yes	yes		regular	http://development.objectvideo.com/index.html
Object/Video	Tool to generate realistic video from simulated cameras in an interactive virtual world.	All types but synthetic	possible	possible	possible	possible	both possible	possible	possible	possible	any	any	http://www.eecs.qmul.ac.uk/~andrea/avss20C/
I-Lids	Scenes of parked vehicle, abandoned package, doorway surveillance and sterile zone.	Airport-like and others	yes	individuals	no	no	both	no	no	yes			http://www-llp.ir.nist.gov/projects/IV2010/
CANDELA	Detection of idle (stationary or non-moving) objects that remain stationary over a certain period of time. The period of time is adjustable. In several types of scenes, idle objects should be detected.	lobby	yes	individuals	no	no	indoor	no	no	yes			http://www-llp.ir.nist.gov/ad/mg/tests/tre
Galwick and I-LIDS MCT airport surveillance video	Part of the datasets used for Trecvid 2010. It doesn't seem to be available for download anymore.	Airport	yes										
AVSS 2010 Multi-Camera Tracking Challenge													http://www-llp.ir.nist.gov/ad/mg/tests/av
Vasor: Outdoor Unimore	Scenes of people entering and leaving a building.	Outdoor courtyard	yes	yes	yes	yes	outdoor	no	no	no	320x240, 10 fps		http://magdabring.unimore.it/veor/video_videos
PETS 2009 with the annotations of the Interactive Graphics Systems Group	Outdoor scenes of individuals and crowds walking.	Outdoor sidewalk	yes	both	no	yes	outdoor	no	no	no			http://www-llp.ir.nist.gov/ad/mg/tests/avss20C/

9 Appendix B: University of Ottawa presentation at the VT4NS 2013 conference

**Canada Border
Services Agency**

**Agence des services
frontaliers du Canada**

Université d'Ottawa | University of Ottawa

Video Analytics study
PROVE-IT




 **uOttawa**


Click View then Header and Footer to change this footer

www.uOttawa.ca

The PROVE-IT (VA) study report



- The report includes:
 - methodology for evaluating VA solutions using sets, mockups, and pilots
 - Development and evaluation of an Interactive Surveillance Event detection system (TRECVID)
 - Survey of Industry and Academic VA solutions & datasets
 - Analysis of the Technology Readiness of some VA components:
 - Unattended/left-behind baggage detection
 - Person tracking (crowded, non-crowded)
 - Person-baggage tagging/association
 - Object removal detection
 - Loitering detection
 - Tail-gating detection
 - Camera tampering detection

 **uOttawa**

Click View then Header and Footer to change this footer

Evaluation of VA systems



- The performance of any detection algorithm can be evaluated from the following measures:
 - TP: number of true positives i.e. the number of good detections;
 - FP: number of false positives, i.e. the number of wrong detections;
 - TN: number of true negatives, i.e. the number of correct nondetection (not always countable);
 - FN: number of false negatives, i.e. the number of missed detections



Click View then Header and Footer to change this footer

Evaluation of VA systems



- The following measures are used to quantify the global performance of a VA system
 - Precision = $TP / (TP + FP)$;
 - Recall (or Sensitivity) = $TP / (TP + FN)$;
 - F1 (or FScore) = $2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$;
- Generally there is a trade-off between precision and recall



Click View then Header and Footer to change this footer

Evaluation of VA systems



- Depends on the business requirements associated with the problem to be solved.
- Operational tolerance should be clearly expressed in terms of
 - acceptable false alarm rate (specified in number of erroneous detections per hour)
 - acceptable misdetection rate (specified as average percentage of undetected events of interest).
- Particular viewing conditions have an important impact on overall performance.



Click View then Header and Footer to change this footer

Video surveillance setups



- **Type 1 (fixed light, person lane):** principle inspection lane (PIL). E.g., passport control point.
- **Type 2 (fixed light, small crowd):** controlled chokepoint. A small number of individuals (1 to 4) at a time following the same direction.
- **Type 3 (fixed light, large crowd):** uncontrolled chokepoint. Four or more individuals at a time following the same direction.
- **Type 4 (variable light, small crowd):** indoor with uncontrolled light (e.g., airports with large windows), or outdoors locations with small crowds (1 to 4 people).
- **Type 5 (variable light, large crowd):** free flow outdoors with four or more individuals.



Click View then Header and Footer to change this footer

Example: loitering detection setups



versus



Click View then Header and Footer to change this footer

Datasets



- Several events and datasets for the evaluation of video analytics technologies
 - PETS
 - AVSS
 - CLEAR
 - NIST TRECVID
 - VACE
 - CAVIAR
 - iLids
 - FRH



Click View then Header and Footer to change this footer

PETS



- Performance Evaluation of Tracking and Surveillance
 - One of the oldest workshop series
- Interested by
 - performance evaluation methodologies for tracking and/or surveillance
 - person counting/density estimation
 - people tracking
 - flow detection/event detection
- Large dataset with ground truth



uOttawa

Click View then Header and Footer to change this footer

i-Lids



- Imagery Library for Intelligent Detection Systems
- UK government's benchmark for Video Analytics (VA) systems
- Dataset available for different event detection scenarios:
 - Parked vehicles
 - Abandoned baggage
 - Doorway surveillance
 - Multi-camera tracking
 - etc.
- Hours of videos mainly from Gatwick airport



uOttawa

Click View then Header and Footer to change this footer

TRECVID



- TRECVID is a workshop series sponsored by the US National Institute of Standards and Technology (NIST)
- Objective: to encourage the development of new technology in information retrieval in video.
- Each year:
 - a series of video retrieval tasks is defined
 - researchers are invited to submit results of their approaches in solving the proposed tasks.



uOttawa

Click View then Header and Footer to change this footer

TRECVID



- Examples of tasks
 - shot boundary detection: detect in a video all the cuts and transitions that separate the different camera shots of the video.
 - story segmentation: given a long video, segment it into sequences of coherent stories, i.e. detects the transitions where the video switches from one topic to another one.
 - video copy detection: from a video query, try to find in a large video database if it exists copies of some segments of this video. The copies might have been altered by some transformations such as cropping, resizing, text overlay.
 - video summarization: from a long video file, create a summary containing the main elements of the original video. Typical summary duration was 30 seconds for a 30 minutes video.
 - multimedia event detection: from a set of 20 predefined events, find a video collection all the occurrences of each of the selected events. Examples of events are birthday party, parade, dog show, etc.



uOttawa

Click View then Header and Footer to change this footer

Interactive surveillance event detection



- interactive surveillance event detection:
 - given a search request for a particular event
 - produce a system that will assist a human operator to identify occurrences of this event in a set of surveillance videos
 - return a list of video segments within the surveillance data files, ranked by likelihood
 - each search for an event by a searcher can take no more than 25 elapsed minutes



Click View then Header and Footer to change this footer



Canada Border
Services Agency

Agence des services
frontaliers du Canada

Université d'Ottawa | University of Ottawa

VIVA Research Lab

Interactive Surveillance Event Detection

uOttawa:

Chris Whiten, Robert Laganière,

Ehsan Fazl-Ersi, Feng Shi

CBSA Science & Engineering Directorate:

Dmitry Gorodnichy, Jean-Philippe Bergeron,

Ehren Choy, David Bissesser

Ecole Polytechnique Montreal:

Guillaume-Alexandre Bilodeau



uOttawa

www.uOttawa.ca

Background



- New interactive SED task
- First participation to SED task
- Limited submission results
 - *Person-runs event detection*
- uOttawa works on automatic the event detection part
- CBSA works on the interactive part



Design objectives



- Problem of high relevance to CBSA
- **To improve computational performance**
- Traditional framework:
 - spatiotemporal features detection
 - Bag of words representation
 - SVM classifier
- Inspiration from MoSIFT (from CMU)
- Inspiration from recent fast image matching techniques
 - Fast feature detector
 - Binary descriptor



Operational need



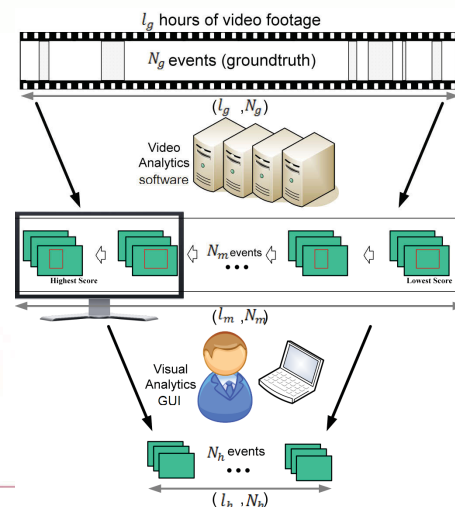
- Surveillance cameras are heavily used by CBSA (in particular, in Airports)
- Two modes of operation:
 - Real-time: eg. to send a traveler to secondary examination
 - Post-event: eg. evidence extraction
- In either mode, the decision - to trigger or not trigger alarm - needs to be made within limited amount of time



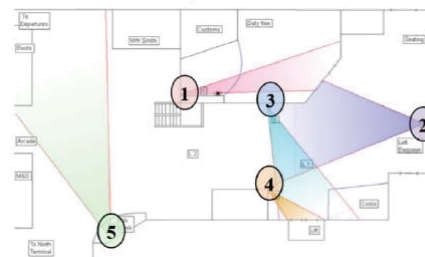
Machine-Human approach



- Current *Video Analytics* algorithms produce lot of false alarms
 - Filtering such amount of false alarms requires efficient *Visual Analytics* tools (GUI) ...
- ... that makes use of humans visual recognition power for fast processing of large quantities of data



TRECVID dataset



Click view then header and footer to change this footer



Event detection by Video Analytics

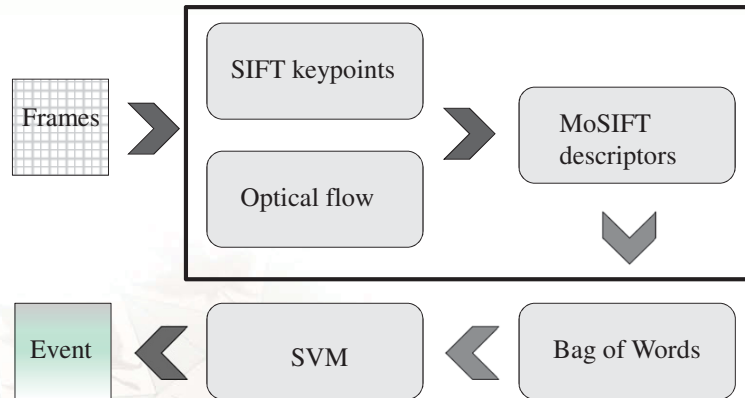


- Most Video Analytics approach are based on space-time points
- Historically, spatiotemporal descriptors have used gradient-based features (SIFT, Histogram of Oriented Gradients, etc..)
 - Slow to detect/compute/match
 - Difficult for the massive scale of surveillance data
- MoSIFT is a good example of such a space-time descriptor



MoSIFT Approach

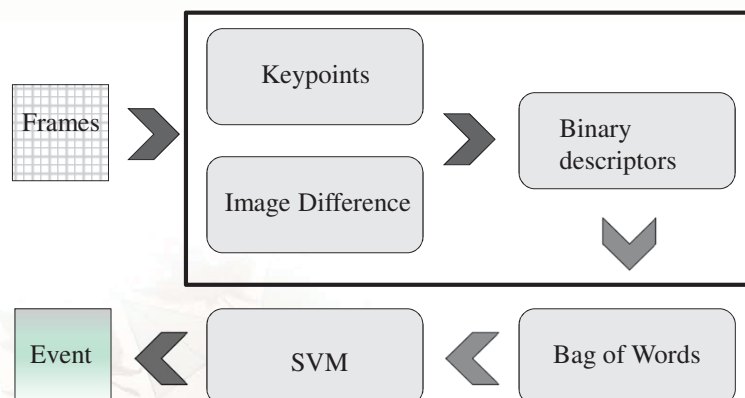
CBSA ASFC



 uOttawa

Our MoFREAK Approach

CBSA ASFC

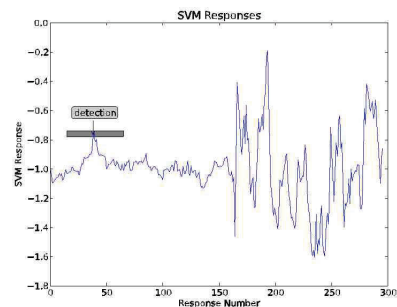


 uOttawa

Automated Event Detection



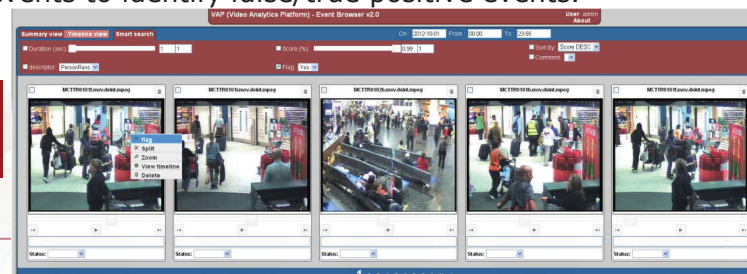
- A video sequence is inputted
- The classifier gives a distribution with many peaks and valleys
- Sufficiently large local maxima = event occurrence



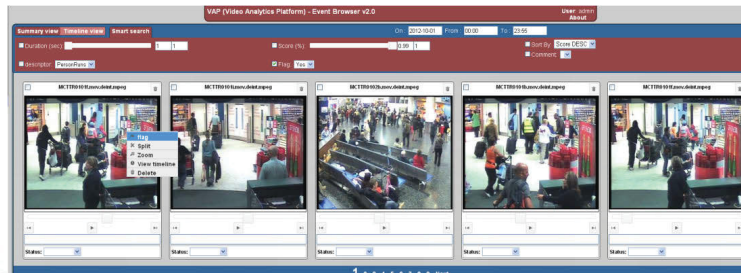
Manually Filtering False Positives



- The event detection system yields many false positives
 - Requires human feedback to know which detected events are legitimate
- Visual analytics system:
 - Events are presented in order of SVM response
 - to allow a user to efficiently navigate detected events to identify false/true positive events.



VAP Browser interface



- Using this visual analytics platform, a human operator is able to process over 600 detected events in a 25 minute time-window (24 events per minute)



Results for the Person-runs event



- There were 107 true events
- We extrated 15 events
 - 6 of those 15 events were deemed to be true events
 - while 9 were deemed false alarms
 - giving us 101 missed detections
- Our rate of false alarms is 0.59027 (RFA)
- Our percentage of missed detections is 0.944 (PMiss).
- The weighted linear combination of the false alarm rate and probability of a missed detection is 0.9469 (DCR).



Click View then Header and Footer to change this footer

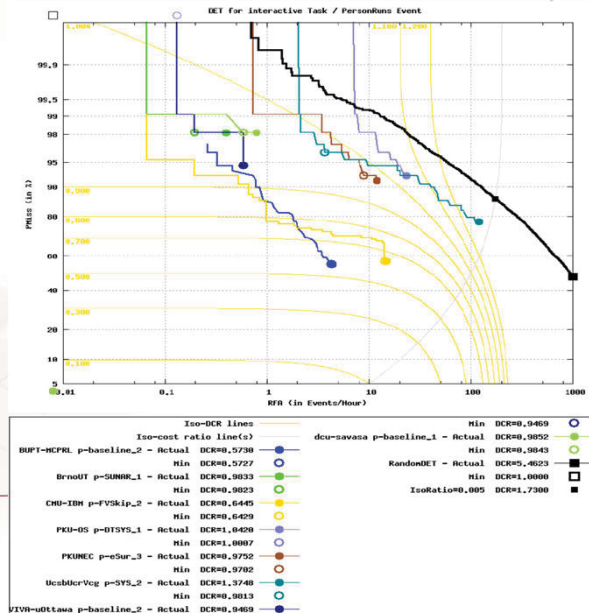
TRECVID submission



- We submit the results for the *person-run* event
 - Events were detected using MoFREAK approach
 - Events were filtered using VAP browser



uOttawa



Person-runs Detections



<http://www.site.uottawa.ca/~laganier/video/runs.avi>



uOttawa

Conclusion



- Using recent advances in binary descriptors, rather than gradient-based descriptors, makes processing surveillance footage much more feasible
 - Currently 3 times faster
- Machine-human approach should however prevail:
 - Video Analytic component allows to detect alarms automatically
 - Visual Analytic interface is critical for efficient filtering of false alarms.



Computational cost



Approach	Running Time (mins)	Accuracy
<i>Fathi and Mori</i> [16]	10,129	90%
<i>Jhuang et al.</i> [17]	1,198	90 %
<i>MoSIFT</i> [6]	449	87 %
<i>MoFREAK</i>	185	90 %

As measured on KTH dataset for action recognition
(600 short sequences)



Click View then Header and Footer to change this footer